Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 1, Maret 2021, pp. 417-425
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Implementasi *Data Science* dalam Ritel *Online*: Analisis *Customer Retention* dan *Clustering Customer* dengan Metode K-Means

Irma Permata Sari
Sistem dan Teknologi Informasi, Fakultas Teknik, Universitas Negeri Jakarta
Kampus A UNJ Gedung L, Jl. Rawamangun Muka Raya, Rawamangun, Jakarta Timur
irmapermatasari@unj.ac.id

Abstract

Data regarding huge sales and purchase transactions are stored electronically in company databases. Leave this data alone so it will not have any impact. Lately, many companies provide promo prices to customers to attract customers. However, the decision is suitable to be taken, regardless of sales growth, to not cause more significant losses. The sales show that has been recorded each year in the sales transaction database. This research focuses on implementing data science at a retail company to analyze sales performance using the cohort analytics method to calculate customer retention and perform clustering customer using the K-Means model. As the results, we can conclude that the company has sales performcean about 37.4%, seen from customer retention and the monthly sales volume within one year. There are three groupings produced, namely ID 0, 1, and 2. Customers with Cluster Label 2 are customers with the highest number of transactions compared to other groups.

Keywords*: data science, retail online, customer retention, clustering customer, K-Means*

Abstrak

Data mengenai transaksi penjualan dan pembelian yang berjumlah sangat besar tersimpan secara elektronik dalam database perusahaan. Data ini apabila dibiarkan begitu saja maka tidak akan memberikan dampak apa-apa. Belakangan ini banyak perusahaan memberikan harga promo yang kepada customer dengan tujuan untuk menarik customer. Namun apakah keputusan tersebut tepat untuk diambil, tanpa mengetahui pertumbuhan penjualan sehingga tidak menimbulkan kerugian yang lebih besar. Performa penjualahan yang telah dicatat setiap tahunnya dalam database transaksi penjualan. Pada penelitian ini akan membahas mengenail implemetasikan data science terhadap suatu perusahaan ritel untuk menganalisis performa penjualan menggunakan metode cohort analytics untuk menghitung customer retention dan melakukan clustering customer menggunakan model K-Means. Dari hasil eksperimen yang dilakukan dapat diketahui bahwa perusahaan mengalami pertumbuhan sebesar 37,4 %, dilihat dari Analisa customer retention dan volume penjualan perbulan dalam kurun waktu satu tahun. Terdapat 3 pengelompokan yang dihasilkan yaitu ID 0, 1 dan 2. Customer dengan cluster label 2 adalah customer yang paling tinggi jumlah transaksinya dibandingkan kelompok lain.

Kata kunci*: data science, retail online, customer retention, clustering customer, K-Means*

1. PENDAHULUAN

Perkembangan perusahaan ritel di Indonesia sangat menarik untuk disimak. Pada tahun 2017 terjadi penurunan kinerja sektor ritel, hal ini ditandai dengan penutupan sejumlah gerai supermarket besar khususnya di Jakarta, seperti Seven Eleven, Lotus, Matahari Departmen Store dan lain-lain [1]. Hal ini disebabkan karena adanya perubahan kondisi perekonomian dan

tren teknologi dalam berbelanja, missal penurunan daya beli masyarakat dan trend di sejumlah kota-kota besar beralih ke pemanfaatan *online shopping*.

Kemudian pada tahun 2020, ritel online berkembang sangat pesat. Hal ini disebabkan pula oleh kondisi pandemic Covid-19, dimana berbelanja online dianggap sebagai solusi di era kebiasaan baru. Toko ritelpun kemudian digantikan oleh toko ritel *online*. Kondisi ini membuat para perusahaan *ritel* harus tetap bertahan dan mengikuti tren teknologi dan pola yang berkembang di masyarakat untuk tetap dapat menjalankan usahanya. Beberapa strategi digunakan yaitu penggunaan marketing digital. misal dengan memanfaatkan *social media* dan *influencer, branding* sebuah produk, dan memberikan potongan harga atau promo pada pelanggan. Banyak sekali perusahaan memberikan harga promo yang kepada *customer* dengan tujuan untuk menarik *customer* agar membeli produk mereka.

Persaingan harga yang begitu ketat, hingga terjadi muncul suatu istilah 'perang harga' dan 'bakar duit' untuk sebuah promo. Dari sisi *customer* tentu hal ini sangat menyenangkan dan mneguntungkan. Namun dari sisi perusahaan keputusan ini harus dianalisis terlebih dahulu secara terukur, agar dapat membuktikan bahwa keputusan yang diambil memiliki pengaruh yang signifikan terkait dengan performa penjualannya. Performa penjualahan yang telah dicatat setiap tahunnya dalam suatu *database*.

Data mengenai transaksi penjualan dan pembelian yang berjumlah besar tersimpan secara elektronik dalam *database* perusahaan dapat dimanfaat untuk menghitung pertumbuhan pelanggan. Pada penelitian ini akan membahas mengenai implemetasikan data science terhadap suatu ritel online untuk menganalisis customer retention dengan menggunakan metode *cohort analytics* dan pengelompokkan pelanggan *(clustering)* dengan algoritma K-Means. Data yang digunakan adalah data dari sebuah perusahaan ritel *online* dengan 541.909 baris dan memiliki 8 kolom. Data ini diambil dari data penjualan perusahaaan tersebut selamat 1 tahun terakhir.

2. METODOLOGI PENELITIAN

2.1. Data

Data adalah kumpulan infomasi baik dalam bentuk terorganisir maupun tidak terorganisir. Kategori data dalam:

a) Data terorgansir *(organized data)* mengacu pada data yang terurut dalam baris / kolom yang secara terstruktur, dimana setiap baris mewakili satu nilai dan kolom mewakili karakteristik nilai tersebut.
b) Data tidak terorganisir *(unorganized data)* mengacu pada jenis data yang bentuknya bebas, biasanya berupa *teks raw audio* / sinyal yang harus dirubah menjadi terstruktur. Missal, ketika anda melihat data di Excel (atau spread sheet lainnya), anda melihat struktur baris/kolom kosong menunggu untuk diorganisir.

2.2. *Data Science*

Dalam arti sempit, *data science* atau ilmuan data adalah seni dan ilmu untuk memeproleh pengetahuan melalui data [2]. Sementara dalam pengertian yang lebih luas, ilmuan data (*data science*) adalah segala hal tentang bagaimana mengambil data, menggunakan data untuk memeproleh pengetahuan, dan kemudian menggunakan pengetahuan tersebut untuk melakukan hal berikut:

a) Membuat keputusan.
b) Memprediksi masa depan.
c) Memmahami masa lalu/sekarang.
d) Menciptakan industri baru /produk baru.

2.3. *Clustering* dengan metode K-Means

Clustering adalah suatu metode yang digunakan untuk pengelompokan sejumlah data menjadi kelompok-kelompok data tertentu *(cluster).* Salah satu algoritma clustering yang paling banyak digunakan adalah K-Means. Algoritma K-Means dapat melakukan pengelompokkan data (clustering) dengan metode partisi dan melakukan pemodelannya tanpa supervisi dan pengelompokan dilakukan kesamaan karateristik-karakteristik [3][4]. Algoritma K-means pada dasarnya bekerja dengan cara *centroid*, dimana sebagian dari banyak komponen diambil untuk kemudian dijadikan *cluster* awal, kemudian penentuan *cluster* ini dipilih secara acak dari populasi data [5].

2.4. Kinerja Perjualan *(Sales performance)*

Kinerja adalah indikator keberhasilan sesorang dalam menyelesaikan pekerjaannya dengan baik dan sesuai standar yang ditetapkan. Kinerja penjualan *(sales performance)* menurut Churchil dkk (1997) dalam [6] adalah peningkatan yang dicapai oleh tenaga penjual *(sales people)* dan divisi penjualan, yang mencerminkan kerja keras dan kerja cerdas untuk mencapai tujuan organisasi. Ferdinan (2002) dalam [6] juga memaparkan bahwa indikator pekerjaan adalah *volume* penjualan, pertumbuhan pelanggan dan pertumbuhan penjualan. Berdasarkan uraian di atas, kinerja tenaga penjualan merupakan ujung tombak dari: proses pemasaran yang dimulai dari penyaluran kredit, pembuatan program promo, kolektibilitas, dan keuntungan yang diperoleh dari bunga dan administrasi.

2.5. Pertumbuhan penjualan *(Sales growth)*

Menurut Kusuma (2009) dalam [7] pertumbuhan penjualan adalah kenaikan penjualan dari tahun ke tahun atau dari waktu ke waktu. Untuk menghitung pertumbuhan penjualan dengan menggunakan *sales growth*.

2.6. *Customer Retention*

Customer retention yaitu suatu kondisi dimana sejumlah *customer* kembali melakukan transaksi pada suatu ritel atau organisasi. Customer

Relationship Management lebih memilih metode mempertahankan pelanggan yang ada untuk mendapatkan keuntungan ketimbang mencari pelanggan baru. Hal ini dikarenakan metode ini ini lebih hemat, bahkan 5 sampai 20 kali lebih hemat biaya dari memperoleh pelanggan yang baru [8].

2.7. *Cohort analitics*

Cohort adalah sekumpulan individu yang masuk kedalam sebuah sistem pada waktu yang sama. *Cohort* juga dapat dikatakan sebagao sekelompok entitas. *Cohort* bisa berupa sekumpulan orang, mobil, pohon, paus, bangunan, kemungkinannya tidak terbatas. *Cohort* biasanya diukur berdasarkan tahun atau periode waktu tertentu untuk dapat dikelompokan ke dalam kelompok tertentu[9]. Sedangkan *cohor*t analisis adalah sebuah analisis yang menjelaskan tentang perbedaan antara dua dimensi temporal lainnya, missal "usia" (waktu masuk ke sebuah sistem) dan "periode" (waktu ketika hasil diukur). Kemudian waktu interval yang menentukan keanggotaan dalam sebuah *cohort* bergantung pada pertimbangan analisis dan kondisi yang ingin diteliti. Ketergantungan linier dari tiga dimensi dalam analisis ini sering menciptakan masalah dalam identifikasi. Namun paling penting dalam pengenalan variabel yang akan diukur harus mengetahui dimensi yang mendasari, misal satu usia, periode, dan kelompok tertentu.

2.8. Metode

Lima (5) tahapan dalam *data science* menurut [2]:

a) Menanyakan sebuah pertanyaan *(Ask the questions)*
   Langkah ini merupakan awal sebelum pekerjaan dimulai. Sangat penting bagi seorang *data scientist* untuk membuat atau memiliki pertanyaan yang kreatif tentang apa yang akan dianalisis, mengingat tidak semua data yang didapatkan berkaitan dengan analisis yang akan dilakukan.
b) Memperoleh data *(Obtaining the data)*
   Setelah memiliki sejumlah pertanyaan dan menentukan apa yang akan analisis, langkah berikutnya adalah mencari data tersebut dari sumber terpercaya. Pada tahapan ini dibutuhkan pengetahuan *database management*, speerti MySQL, PostgreSQL, MongoDB dan lain-lain. Setelah memperoleh data *(data gathering)* kita dapat melakukan sejumlah *query* atau *scrubbing the data*. Proses ini bertujuan untuk membersihkan dan mem-*filter* data dari *noise*, duplikasi data, dan *missing value*. Sebab apabila terdapat data yang tidak relevan maka hasil analisis bisa jadi salah.
c) Mengeksplorasi data *(Exploring the data)*
   Melakukan eksplorasi terhadap data, memanipulasi data, melakukan data transformasi guna untuk mempersiapkan proses pemodelan data.
d) Memodelkan data *(Modeling)*
   Pemodelan data dapat menggunakan model *statistics* dan *machine learning*. Pada tahapan ini juga dilakukan proses pemodelan dan mengukur keefektifannya.

e) Mengkomunikasikan dan memvisualisasikan hasil *(Communicating and visualizing the results)*
Tahapan berikutnya menyimpulkan hasil, kemudian mengkomunikasinnya dalam visualisasi yang baik.

Hampir dengan tahapan diatas, menurut [10] terdapat 6 tahapan dalam *data science* yaitu:

a) *Data Gathering, Preparation, and Exploration.*
b) *Data Representation and Transformation.*
c) *Computing with Data.*
d) *Data Modelling.*
e) *Data Visualization and Presentation.*
f) *Science about Data Science.*

3. HASIL DAN PEMBAHASAN

Penelitian ini merujuk metode yang 5 tahapan dalam *data science* yang dideskripsikan oleh Ozdemir [2]. Untuk lebih detail dapat dilihat pada sub bahasan berikut ini:

3.1. *Ask the questions*

Fokus dari penelitian ini adalah analisis terhadap kinerja penjualan. Jadi beberapa pertanyaan yang harus dijawab oleh data adalah terkait dengan kenaikan penjualan dari dari waktu ke waktu. Dari penelitian ini menggunakan jumlah transaksi perbulan, apakah cendeerung naik atau turun. Kemudian dilihat pula kencerungan pertumbuhan pelangganya, apakah bertumbuh atau tidak melalui analisa *customer retention* dengan metode *cohort*.

3.2. *Obtaining Data*

Data yang digunakan dalam penelitian ini data publik yang diambil dari pangkalan data Kaggle. Proses analisis dilakukan dengan menggunakan *tools* Google Colaboratory. Sebelum masuk ke *exploring data*, beberapa tahapan yang dilakukan adalah:

a) Mengatur koneksi data dan *import* data ke Google Colaboratory

Pengaturan koneksi dari Google Drive ke Google Colaboratory dilakukan dengan menggunakan link *email* Gmail yang terdaftar. Detail *dataset* yang digunakan dapat dilihat pada Gambar 1. Pada Gambar 1 dapat diketahui bahwa *dataset* yang digunakan pada penelitian ini berumlah 541.909 baris dan memiliki 8 kolom, yang terdiri atas: *Invoice Number, Stock Code, Description, Quantity, Invoice date, Unit Price, Customer ID, dan Country.*

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 1, Maret 2021, pp. 417-425
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	09-12-2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	09-12-2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	09-12-2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	09-12-2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	09-12-2011 12:50	4.95	12680.0	France

541909 rows × 8 columns

Gambar 1. Contoh dataset penjualan ritel *online*

b) *Data understanding*

Data understanding dilakukan dengan menggunakan fungsi berikut:

1) Distribusi data

Untuk memahami distribusi ke data-data yang menjadi *concern* dapat dilakukan dengan sintaks data.describe(). Disini yang tampilkan adalah data berupa numerik.

2) Tipe data

Sintaks data.info()ini digunakan untuk memahami tipe data dari setiap kolom.

c) *Data Clean*

1) Mengecek *missing value*

Pengecekan nilai yang kosong dapat dilakukan dengan sintaks:

```
data.isnull().sum()
```

2) Me-*remove missing value*

Salah satu cara untuk mengatasi *missing value* adalah dengan me*remove* data yang kosong tersebut dengan menggunakan sintaks:

```
data = data.dropna()
data.shape
```

3) Mengecek data duplikat

Pengecekan data duplikasi juga merupakan bagian dari data *cleaninng*. Proses ini dapat dilakukan dengan menggunkan sintaks berikut:

```
data.duplicated().sum()
```

4) Menghapus data yang duplikat

Setelah mengetahui data yang terdapat duplikasi, langkat berikutnya adalah menghapus data duplikasi tersebut, dengan menggunakan sintaks:

```
data.drop_duplicates(keep='first', inplace=True)
data.shape
```

d) *Data manipulation*

Data manipulasi dilakukan dengan menggunakan sintaks berikut:

```
retail = data.copy()
```

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 1, Maret 2021, pp. 417-425
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

3.3. *Exploring Data*

Setelah *dataset* bersih dari *missing value* dan duplikasi maka langkah berikutnya adalah proses *exploring data*. Beberapa *exploring* yang dilakukan adalah*:*

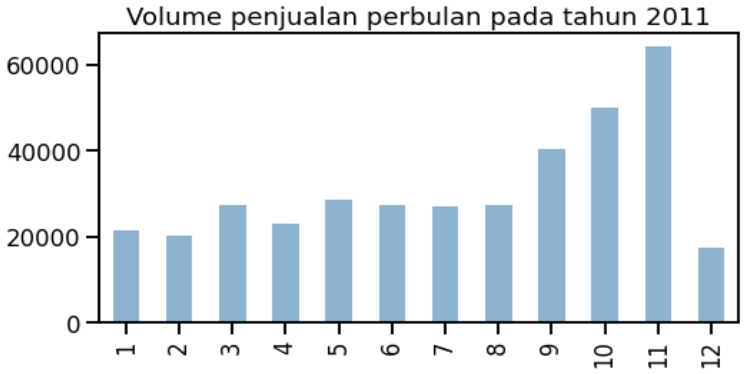
a) Analisa pendapatan per-bulan

Analisis terhadap *volume* penjualan dapat dilakukan dengan sintaks berikut, seperti yang terlihat pada Gambar 2.

```
# volume of transactions per month
plt.figure(figsize=(8,4))
retail[retail.InvoiceDate.dt.year==2011].InvoiceDate.dt.month.val
ue_counts(sort=False).plot(kind='bar')
plt.title("Volume penjualan perbulan pada tahun 2011")
plt.show()
```

Gambar 2. Sintaks untuk menghitung transaksi perbulan

Hasilnya perhitungan *volume* penjulan pada tahun 2011 terlihat pada Gambar 1.



Gambar 3. Visualisasi *volume* penjualan

Tampak pada Gambar 3 bahwa *volume* penjulan tertinggi terjadi pada bulan ke-11 atau November dan terendah pada bulan ke-12 atau Desember.

b) Mengidentifkasi pertumbuhan pelanggan

Analisa untuk mengidentifikasi pelanggan ini dapat dilakukan dengan menganalisis *customer retention*. Untuk dapat mengetahui *customer retention* dapat menggunakan sintaks pada Gambar 4.

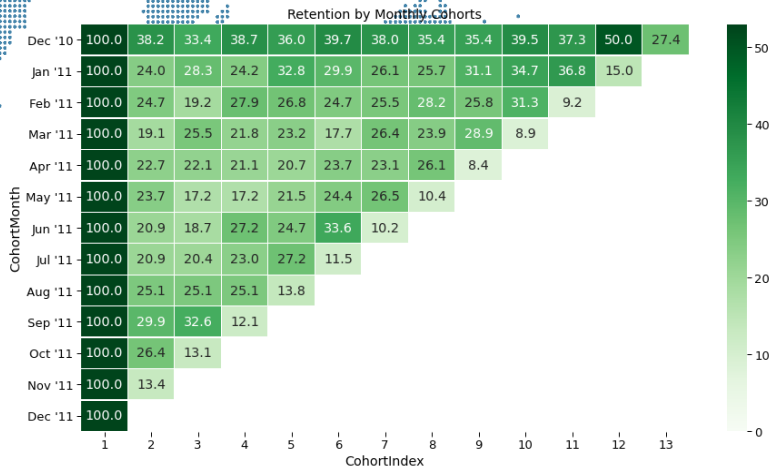
```
#count monthly active customers from each cohort
grouping = retail.groupby(['CohortMonth', 'CohortIndex'])
cohort_data = grouping['CustomerID'].apply(pd.Series.nunique
)
cohort_data = cohort_data.reset_index()
cohort_counts = cohort_data.pivot(index='CohortMonth', colum
ns = 'CohortIndex', values='CustomerID')

#Customer retention
```

Gambar 4. Sintaks program untuk *customer retention*

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 1, Maret 2021, pp. 417-425
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Dari hitungan di gambar 4 dapat dibuat visualisasi sebagaimana terlihat pada Gambar 5.



Gambar 5. Visualisasi *customer retention*

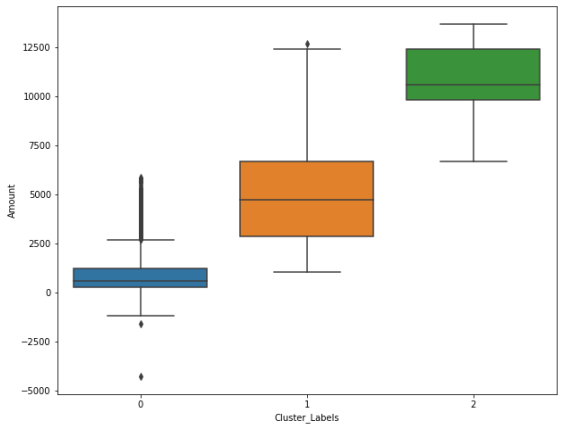
Berdasarkan Gambar 5 dapat diketahui bahwa pertumbuhan pelanggan mengalami penurunan sekitar 27.4 – 50 %, artinya tidak mencapai persentase yang ideal seperti pada bulan pertama atau naik pada tahun 2011.

3.4. *Modelling*

Pada tahapan *modelling* dibangun dengan model K-Means Clustering. Untuk mencari berapa jumlah kluster terbaik dilakukan dengan metode Elbow Criterion. Metode ini merupakan metode yang paling umum digunakan dalam studi kasus *clustering*. Hasil pemodelan dan evaluasi eksperimen ini dapat diketahui hasil pengelompokkannya di sub bab 3.5 berikut ini.

3.5. *Communicating and visualizing*

Hasil visualisasi antara *cluster label* dengan *frequency* (*amount*) penjualan dapat dilihat pada Gambar 6.



Gambar 6. *Clustering Label* dan *Frequency*

Dari Gambar 6 dapat diketahuai bahwa *customer* dengan Cluster Label 2 adalah pelanggan dengan jumlah transaksi yang tinggi dibandingkan dengan pelanggan lainnya. *Customer* dengan Cluster Label 2 ini merupakan *customer* yang sering melakukan transaksi pembelian atau pelanggan tetap. Sedangkan pelanggan dengan Cluster Labels 0 bukanlah pembeli baru dan karenanya tidak terlalu penting dari sudut pandang bisnis.

4. SIMPULAN

Dari data transaksi penjualan yang ada dalam *database* perusahaan tersebut, dapat diketahui bahwa perusahaan tersebut mengalami pertumbuhan pelanggan rata-rata sekitar 37.42% sepanjang tahun 2011, volume penjualan tertinggi terjadi pada bulan November, serta dari 3 pemodelan kluster pelanggan, dinilai bahwa kluster pelanggan dengan label 2 adalah pelanggan tetap dan memiliki jumlah transaksi yang paling besar. Dapat disarankan kepada peneliti berikutnya agar dapat melakukan riset terdapat keputusan yang lebih mendetail untuk pemberian promo dan menganalisis terhadap produk yang paling laku terjual.

DAFTAR PUSTAKA

[1] N. S. Utami, "Analisa kinerja sektor ritel indonesia," vol. 1, no. 1, pp. 43–48, 2018.
[2] S. Ozdemir, *Principles of Data Science*. 2016.
[3] Mardalius, "Pengelompokan Data Penjualan Aksesoris Menggunakan Algoritma K-Means," vol. IV, no. 2, pp. 401–411, 2018.
[4] S. P. Tamba and F. T. F. Kesuma, "Penerapan Data Mining untuk Menentukan Penjualan Sparepart Toyota dengan Metode K-Means Clustering," vol. 2, no. 2, 2019.
[5] A. Ali, "Klasterisasi Data Rekam Medis Pasien Menggunakan Metode K-Means Clustering di Rumah Sakit Anwar Medika Balong Bendo Sidoarjo," vol. 19, no. 1, 2019.
[6] R. M. Hapsari and F. Indriani, "Quality Analysis Of Support Marketing And Quality Of Sales," vol. XVI, no. 3, pp. 145–166, 2017.
[7] H. R. Valentina, "Pengaruh Struktur Modal, Risiko Bisnis Dan Pertumbuhan Penjualan Terhadap Kinerja Keuangan Pada Perusahaan Real Estate And Property Yang Terdaftar di Bursa Efek Indonesia Tahun 2010-2014 By:," vol. 4, no. 2, 2014.
[8] S. F. Sabbeh, "Machine-Learning Techniques for Customer Retention : A Comparative Study," vol. 9, no. 2, pp. 273–281, 2018.
[9] A. Bell, "Age Period Cohort analysis : A review of what we should and shouldn ' t do," no. September 2019, 2020.
[10] D. Donoho, F. Group, D. Donoho, and D. Donoho, "50 Years of Data Science," vol. 8600, 2017, doi: 10.1080/10618600.2017.1384734.